BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1928–1929
doi:10.1093/bioinformatics/btl268

Genetics and population analysis

SNPStats: a web tool for the analysis of association studies

Xavier Solé[1], Elisabet Guinó[1], Joan Valls[1,2], Raquel Iniesta[1] and Víctor Moreno[1,2,*]

[1]Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry, L'Hospitalet, Barcelona, Spain and [2]Autonomous University of Barcelona, Laboratory of Biostatistics and Epidemiology, Bellaterra, Barcelona, Spain

Received on March 6, 2006; revised on May 16, 2006; accepted on May 18, 2006
Advance Access publication May 23, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: A web-based application has been designed from a genetic epidemiology point of view to analyze association studies. Main capabilities include descriptive analysis, test for Hardy–Weinberg equilibrium and linkage disequilibrium. Analysis of association is based on linear or logistic regression according to the response variable (quantitative or binary disease status, respectively). Analysis of single SNPs: multiple inheritance models (co-dominant, dominant, recessive, over-dominant and log-additive), and analysis of interactions (gene–gene or gene–environment). Analysis of multiple SNPs: haplotype frequency estimation, analysis of association of haplotypes with the response, including analysis of interactions.

Availability: http://bioinfo.iconcologia.net/SNPstats. Source code for local installation is available under GNU license.

Contact: v.moreno@iconcologia.net

Supplementary Information: Figures with a sample run are available on *Bioinformatics* online. A detailed online tutorial is available within the application.

The analysis of association between genetic polymorphisms and diseases allows identifying susceptibility genes (Cordell and Clayton, 2005). The proper analysis of these studies can be performed with general purpose statistical packages, but the researcher usually needs the assistance of additional software to perform specific analysis, like haplotype estimation, and results from different packages are difficult to integrate.

We present a free web-based tool to help researchers in the analysis of association studies based on SNPs or biallelic markers. Both the selection of analysis and the output have been designed from a genetic epidemiology perspective. This application can also be used for learning purposes. We have written (in Spanish) an analysis guide with detailed explanations (Iniesta *et al*., 2005). A similar extensive help in English can also be found on the website.

The software is used following three steps, with the possibility of performing multiple analyses in one session. The steps are as follows.

(1) *Data entry*. Raw data in tabular form can be pasted in a window or uploaded from a text file. Variables can be named and the user can choose the field delimiter and the missing value code (Supplementary Figure 1). SNPs should be coded as genotypes with each allele separated by a slash (e.g. 'T/T', 'T/C', 'C/C').

(2) *Data processing*. A list with the variables read by the application is presented with an initial suggestion about the type: quantitative, categorical or SNP, which can be modified (Supplementary Figure 2). The user is prompted to select those needed for the analysis and to specify which one is the response, which may be binary (disease status) or quantitative. For categorical variables, including SNPs, the user can reorder the categories. The first one will be treated as reference category in the analysis. The application assumes that the main interest is the analysis of the SNPs in relation to the response. Other variables selected with type quantitative or categorical will be added to the regression models for analysis as covariates and treated as potential confounders.

(3) *Analyses customization*. The third step requests the selection of the desired statistical analyses that will be described later in this article (Supplementary Figure 3).

Regarding the statistical analysis, the association with disease is modeled depending on the response variable. If binary, the application assumes an unmatched case–control design and unconditional logistic regression models are used. If the response is quantitative, then a unique population is assumed and linear regression models are used to assess the proportion of variation in the response explained by the SNPs.

The association for each SNP is analyzed in turn and adjusted for the selected covariates. If more than one SNP are selected, then the application assumes that haplotype analysis is appropriate. Haplotype frequencies are estimated using the implementation of the EM algorithm coded into the *haplo.stats* package (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm). Association between haplotypes and disease appropriately accounts for the uncertainty in the estimation of haplotypes for individuals with multiple heterozygous when phase is unknown or when missing values are present (Schaid *et al*., 2002). Individuals with missing values in the response, in all SNPs or in any covariate are excluded from analysis.

The software main page can be found online at http://bioinfo.iconcologia.net/SNPstats. The application uses PHP server programming language to build the input forms, upload data, call the statistical analysis procedures and process the output. The statistical analyses are performed in a batch call to the R package (R Development Core Team, 2005, http://www.R-project.org). The contributed packages *genetics* (Warnes and Leisch, 2005) and *haplo.stats* (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm) are called to perform some of the analysis. Anonymous use is guaranteed and data are

*To whom correspondence should be addressed.

treated as confidential. Source code for local installation (Linux and Windows) is also available under GNU license.

SNPStats returns a complete set of results for the analysis, covering from the descriptive statistics to the haplotype analysis. The descriptive statistics returned are the absolute frequencies and proportions for categorical variables, and mean, standard deviation and a list of percentiles for the quantitative ones. Always the total valid sample size and the count of missing values are displayed (Supplementary Figure 4).

Each SNP is described as allele and genotype frequencies. An exact test for Hardy–Weinberg equilibrium is performed (Supplementary Figure 5). When the response variable is binary, these statistics can be displayed by each response group. The user usually will be interested in checking Hardy–Weinberg equilibrium in the control population.

The analysis of association for each SNP can be performed both for quantitative or binary response variables. For binary responses, the logistic regression analysis is summarized with genotype frequencies, proportions, odds ratios (OR) and 95% confidence intervals (CI) (Fig. 1). For quantitative responses, linear regression is summarized by means, standard errors, mean differences respect to a reference category and 95% CI of the differences.

SNPStats can also perform analyses of interactions. For simplicity, models with only one pair of variables interacting can be selected at a time. Three summary tables are shown (Supplementary Figure 7). The first one is the cross-classification that uses a common reference category for both interacting variables. ORs or mean differences are estimated, together with 95% CI, for all other combinations. Next tables use the margins as reference category and estimate ORs or mean differences of one variable nested within the other one. A global test for interaction is performed, as well as a test for the interaction in the linear trend of the nested variable. This assumes that the nested variable is ordinal and tests for different trend among categories. This test might be more sensitive than the global one due to the reduction in degrees of freedom.

When more than one SNP is included in the analysis, SNPStats offers the possibility of performing linkage disequilibrium (LD) and haplotype analysis. For LD, matrices with selected statistics (D, D′, Pearson's r and associated P-values) are shown. (Supplementary Figure 8).

In the analysis of haplotypes, descriptive statistics show the estimated relative frequency for each haplotype (Supplementary Figure 9). Cumulative frequencies are also shown to help in the selection of the threshold cut point to group rare haplotypes for further analysis. The association analysis of haplotypes is similar to that of genotypes in that either logistic regression results are shown as OR and 95% CI or linear regression results with differences in means and 95% CI. The most frequent haplotype is automatically selected as the reference category and rare haplotypes are pooled together in a group. The analysis of haplotypes assumes a log-additive model by default, but dominant and recessive models are available as alternative choices.

When haplotypes are selected for interaction tables similar to the genotype interaction ones are shown, replacing the genotypes by haplotypes (Supplementary Figure 10). This analysis of interactions and presentation of the results is unique to the available alternatives explored and is an important contribution to the analysis of genetic epidemiology studies, often focused on testing for gene–environment interactions (Lake *et al*., 2003).

As a limitation, we are aware that the selection of the available analysis has been done for the most frequent profile but might not be adequate in some instances. We plan to implement in future versions more response types: survival data for studies of prognosis, multinomial data for categorical responses with more than two categories and paired designs (matched case–control or nested case–control).

SNP1 association with response STATUS (n=685, adjusted by SEX+AGE)

Model	Genotype	STATUS=0-Control	STATUS=1-Case	OR (95% CI)	P-value	AIC	BIC
Codominant	C/C	129 (39.9%)	158 (43.6%)	1.00			
	C/G	152 (47.1%)	150 (41.4%)	0.80 (0.58-1.10)	0.29	949.3	972
	G/G	42 (13%)	54 (14.9%)	1.06 (0.66-1.69)			
Dominant	C/C	129 (39.9%)	158 (43.6%)	1.00	0.31	948.7	966.9
	C/G-G/G	194 (60.1%)	204 (56.4%)	0.85 (0.63-1.16)			
Recessive	C/C-C/G	281 (87%)	308 (85.1%)	1.00	0.44	949.2	967.3
	G/G	42 (13%)	54 (14.9%)	1.19 (0.77-1.84)			
Overdominant	C/C-G/G	171 (52.9%)	212 (58.6%)	1.00	0.12	947.4	965.5
	C/G	152 (47.1%)	150 (41.4%)	0.79 (0.58-1.07)			
Log-additive	---	---	---	0.96 (0.78-1.20)	0.74	949.7	967.8

Fig. 1. Sample output of association models for binary response. Five inheritance models are fitted, which correspond to different parameterizations or groupings of the genotypes. Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are calculated to help the user in the selection of the best model for a specific SNP.

ACKNOWLEDGEMENTS

Funding support from the Spanish Instituto de Salud Carlos III (networks of centres RCESP C03/09 and RTICCC C03/10). Funding to pay the Open Access publication charges for this article was provided by Instituto de Salud Carlos III (FIS 03/0114).

Conflict of Interest: none declared.

REFERENCES

Cordell,H.J. and Clayton,D.G. (2005) Genetic association studies. *Lancet*, **366**, 1121–1131.

Iniesta,R. *et al*. (2005) Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos. *Gac. Sanit*., **19**, 333–341.

Lake,S. *et al*. (2003) Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous. *Human Heredity*, **55**, 56–65.

R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Schaid,D.J. *et al*. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet*., **70**, 425–434.

Sinnwell,J.P. and Schaid,D.J. (2005) , haplo.stats: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.2.2.

Warnes,G. and Leisch,F. (2005) Genetics: Population Genetics. R package version 1.2.0.